The Process of Knowledge Discovery from Large Pharmacokinetic Data Sets

Ene I. Ette, PhD, FCCP, FCP, Paul Williams, PharmD, He Sun, PhD, Emmanuel Fadiran, PhD, Funmilayo O. Ajayi, PhD, FCP, and Leonard C. Onyiah, PhD, FRSS

The advent of statistical software with powerful graphical and modeling capabilities has revolutionized the manner in which pharmacokinetic and pharmacodynamic analyses are performed. Knowledge discovery from a large (population) pharmacokinetic data set incorporates all steps taken from data assembly to the development of a population pharmacokinetic model and the communication of the results thereof. The process can be formalized into a number of steps: (1) creation of a data set for pharmacokinetic knowledge discovery, (2) data quality analysis, (3) data structure analysis (exploratory examination of raw data), (4) determination of the basic pharmacokinetic model that best describes the data and generating post hoc empiric individual Bayesian parameter estimates, (5) the search for patterns and relationships between parameters and parameters and covariates by visualization, (6) the use of modern statistical modeling techniques for data structure revelation and covariate selection, (7) consolidation of the discovered knowledge into irreducible form (i.e., developing a population pharmacokinetic model), (8) the determination of model robustness (determination of the reliability of model parameter estimates), and (9) the communication and integration of the discovered pharmacokinetic knowledge. This process is discussed, and a motivating example is presented. The use of modern graphical, modeling, and statistical techniques for knowledge discovery from large pharmacokinetic data sets has given the data analyst the freedom to choose statistical methodology appropriate to the problem at hand with the maximization of information extraction, rather than on the basis of mathematical/statistical tractability.

Journal of Clinical Pharmacology, 2001;41:25-34

©2001 the American College of Clinical Pharmacology

Population pharmacokinetics is the study of the sources and correlates of variability in plasma drug concentrations between individuals who are preferably the target patient population receiving clinically relevant doses of a drug of interest.[1] Certain patient pathophysiological features, such as body weight, excretory and metabolic functions, and the presence of other therapies, can regularly alter dose-concentration relationships. For example, renal failure usually causes steady-state drug concentrations to be greater than those in patients with normal renal function receiving the same dosage of a drug eliminated mostly by the kidney. Population pharmacokinetics seeks to identify the measurable pathophysiologic factors that cause changes in the dose-concentration relationship and the extent of these changes so that if such changes are associated with clinically significant shifts in the therapeutic index (i.e., safety margin), dosage can be appropriately modified.

For years, most pharmacokineticists have developed population pharmacokinetic models for drugs without a careful exploratory examination and modeling of the multivariate large (population) pharmacokinetic data set. Population pharmacokinetic modeling in itself is a process of knowledge discovery from a population pharmacokinetic data set. Knowledge discovery in this sense is broader than the narrow definition of knowledge discovery as being the search for relationships and global patterns that exist in large da-

From Vertex Pharmaceuticals, Inc., Cambridge, Massachusetts (Dr. Ette); School of Pharmacy and Allied Health Professions, University of the Pacific, Stockton, California (Dr. Williams); Office of Clinical Pharmacology and Biopharmaceutics, Center for Drug Evaluation and Research, Food and Drug Administration (FDA), Rockville, Maryland (Dr. Sun, Dr. Fadiran, Dr. Ajayi); and Department of Statistics, St. Cloud State University, St. Cloud, Minnesota (Dr. Onyiah). The opinions expressed are those of the authors and not those of the U.S. Food and Drug Administration. Submitted for publication December 2, 1999; revised version accepted July 6, 2000. Address for reprints: Ene I. Ette, PhD, Vertex Pharmaceuticals, Inc., 130 Waverly Street, Cambridge, MA 02139.

PM3000456974

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

tabases but are "hidden" among the vast amounts of data.[2] Knowledge discovery from a population pharmacokinetic data set incorporates all steps taken from data assembly to the development of a population pharmacokinetic model.

The advent of modern statistical software such as S-Plus™ with its powerful graphical and modeling capabilities has revolutionized the manner in which pharmacokinetic and pharmacodynamic analyses are performed. Careful exploratory examination and modeling of data (i.e., exploratory data analysis [EDA]) are fundamental steps in the development of a population pharmacokinetic model.[4] The EDA is an essential precursor to and interwoven with the nonlinear mixed-effects modeling step in that it can reveal information patterns and relationships or special structure in the data. Such information can expedite the selection of appropriate modeling steps with a reduction in the time spent in developing population pharmacokinetic models. Thus, the knowledge discovery process can be facilitated with EDA.

Most classical statistical procedures or types of software are based, either implicitly or explicitly, on assumptions about data, and the validity of the analyses depends on the validity of the assumptions. Good data-visualization tools provide powerful diagnostic mechanisms for confirming assumptions or, when the assumptions are not met, for suggesting corrective action.[4]

One of the most difficult tasks for a pharmacometrician/pharmacokineticist is to convey findings from pharmacostatistical analyses to clinicians and other members of the drug development team. Failure to communicate these findings successfully puts at risk all his or her data analysis efforts, irrespective of its quality. The use of high-quality graphics can effectively enhance the communication of pharmacometric knowledge to a medical research team. Graphics, in particular, are essential for conveying relations and trends in an informal and simplified visual form.

The remainder of the article is divided into the following sections: a process of knowledge discovery from a large (population) pharmacokinetic data set, a motivating example of the implementation of the pharmacokinetic knowledge discovery process, discussion, and concluding remarks.

THE PHARMACOKINETIC KNOWLEDGE DISCOVERY PROCESS

The purpose of data analysis and interpretation, in general, is to find out, among other things, meaningful patterns and relationships between variables under consideration. It is analysis that transforms data into useful information.[5] In other words, the purpose of a population pharmacokinetic data analysis is to extract knowledge available in the population pharmacokinetic data set. Sufficient knowledge of the problem investigated and knowledge of mathematics and statistics are essential criteria for a successful pharmacometric evaluation of a population pharmacokinetic data set. Significant untapped knowledge often lies hidden in a population pharmacokinetic data set. Knowledge discovery is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[2] The challenge of knowledge discovery in a population pharmacokinetic data set is to make effective use of the data set to discover the untapped knowledge that lies hidden therein. Specifically, it is the task of implementing (and developing) methodologies that can lead to the discovery of interesting, useful patterns and relationships that can be used (1) to support mission-critical decision making in drug development or drug therapy, (2) for predictions, (3) for the explanation of variability and making dosage recommendations, or (4) for relating pharmacokinetics to pharmacodynamics, efficacy, and safety. These constitute some of the objectives of population modeling.

Knowledge discovery is an emerging interdisciplinary research field that lives at the intersection of computer science (database, artificial intelligence, graphics, and visualization), statistics, and several application domains such as clinical pharmacology in general and pharmacometrics in particular. Knowledge discovery in a large pharmacokinetic data set is a process that can be formalized into a number of steps. Briefly, these steps are as follows:

(1) creating a data set on which knowledge will be performed;
(2) cleaning and processing the data (i.e., data quality analysis);[6,7]
(3) data structure analysis—exploratory examination of raw data (concentrations and covariates) for hidden structure and the reduction of the dimensionality of the covariate vector;
(4) determining the basic pharmacokinetic model that best describes the data and generating post hoc empiric individual Bayesian parameter estimates;
(5) searching for patterns and relationships between parameters, and parameters and covariates through graphical displays and visualization;
(6) using modern statistical modeling techniques such as multiple linear regression (MLR), generalized additive modeling (GAM),[8] and tree-based modeling (TBM) to

PM3000456975

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

reveal structure in the data and initially select explanatory covariates;
(7) consolidating the discovered knowledge in (6) into irreducible form (i.e., developing a population pharmacokinetic model using the nonlinear mixed-effects modeling approach);
(8) determining model robustness through sensitivity analysis, examination of parametric/nonparametric standard errors, or stability, determination, and predictive performance; and
(9) communicating and integrating the discovered pharmacokinetic knowledge.

Generally, not much attention has been given to data ıality analysis or structure revelation in a population ıarmacokinetic data set that could provide a link beeen the data set and the analytical path chosen for ıarmacokinetic knowledge discovery. A lack of these alyses can result in the production of biased popula-ın pharmacokinetic parameter estimates. To avoid as caused by a nonsystematic data analysis, the above uctured approach has been proposed.

Aspects of data quality that need to be analyzed are rrectness and completeness. Correctness of concention-time data relative to dosing history and coriate information can be checked by comparing the cords in the population pharmacokinetic data set th information in the case report forms, and this can done by using a sample of the records. Completeness population pharmacokinetic (and pharmacodymic) data records is a property that cannot be satisd due to many reasons such as omissions when reding or inputting data into a clinical database or the ılfunctioning of medical equipment.[9] To prepare ta for population pharmacokinetic knowledge disvery, some imputation of data may be done, and re are different procedures available for handling ssing data.[6,10,11]

Data structure analysis is the examination of the raw a for "hidden" structure, outliers, or leverage obserions. This is further repeated during the exploratory deling (and nonlinear mixed-effects modeling) ps using case deletion diagnostics.[4] This type of ılysis is important since outliers or leverage observans may occur in a population pharmacokinetic data

The knowledge discovery basis of population pharcokinetic modeling permits the generation of hyhesis from the relationship discovered during data ıcture analysis, which can be further tested in the ılinear mixed-effects modeling step. It can also sugt a testable hypothesis that can be independently ted via traditional confirmatory experiments and ılysis.

A MOTIVATING EXAMPLE

Data from 138 subjects (comprising both sexes with ages ranging from 5 to 81 years and weighing between 17 and 96 kg) who were administered different doses of a drug under development were pooled from six studies for the development of a population pharmacokinetic model (see Ette and Ludden[4]).

A structured approach to the development of a population pharmacokinetic model proposed in the preceding section was used.[4] This was done to uncover "hidden" structure in the multivariable data that informed the analysis path taken to develop a population pharmacokinetic model. After ensuring data quality, the steps taken to develop the population pharmacokinetic model were the following:

1. data structure analysis (i.e., exploratory examination of raw concentration data, distributions, and correlations between covariates);
2. determination of a basic pharmacokinetic model using the NONMEM program and generation of empiric Bayesian individual parameter estimates obtained for subsequent use in searching for patterns and relationships between parameters and covariates;
3. use of empiric Bayesian estimates generated in (2) with covariate data in MLR with case deletion diagnostics, GAM, and TBM for the initial selection of covariates and further revelation of "hidden" structure in the data;
4. consolidating the knowledge discovered in step (3) into an irreducible (final) form through nonlinear mixed-effects modeling to develop the population pharmacokinetic model;
5. the determination of the reliability of parameter estimates using the jackknife technique; and
6. communication of the discovered knowledge.

Exploratory Examination of Raw Concentration Data, Distributions, and Correlations between Covariates

The concentration data from each of the studies were plotted and examined independently of each other for clustering of observations using a "spaghetti" (profile) plot and later pooled and the plot repeated. Figure 1 shows a clustering of observations into two groups in the terminal elimination phase of the plasma concentration-time profile.

A slight bimodality in creatinine clearance (CRCL) is observed in Figure 2. This was confirmed with formal testing using Shapiro-Wilk's test.[12] Figure 2 summarizes the distributions of continuous covariates in the population pharmacokinetic data set.[4]

PM3000456976

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

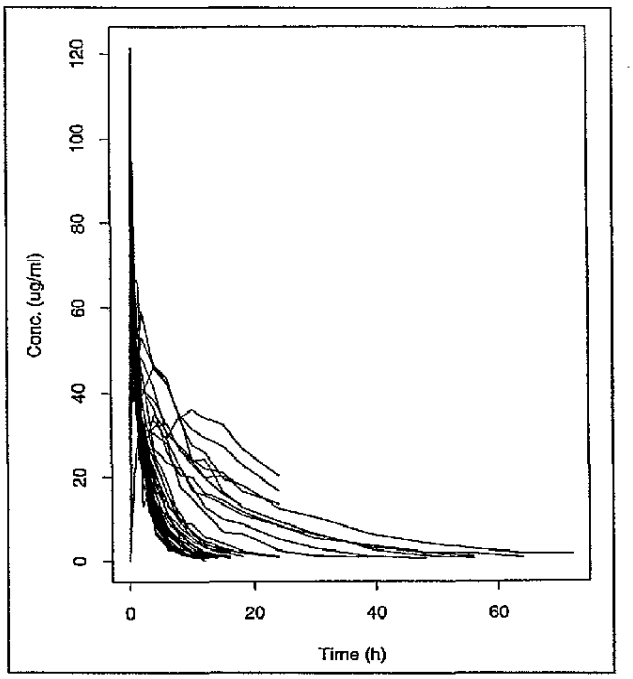


Figure 1. "Spaghetti" plot of individual plasma profiles (in clusters) from a population pharmacokinetic data set of a drug under development.

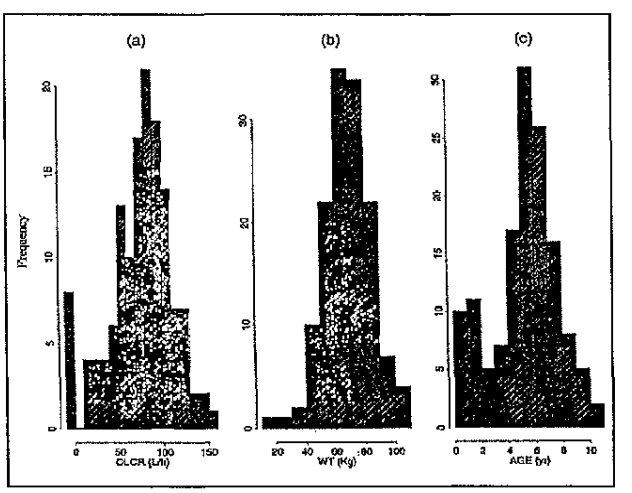


Figure 2. Frequency distribution of demographic variables: (a) creatinine clearance (CRCL), (b) weight (WT), and (c) age. (Extracted from Ette and Ludden[5])

The correlations of weight and age, CRCL and age/weight were formally tested. Although weight was significantly correlated with age ($p < 0.0001$), the correlation coefficient was 0.34. Similarly, creatinine cl ance was significantly correlated with age ($p < 0.00$ but the correlation coefficient was 0.29. Creatin clearance was not correlated with weight ($r = 0.0$ Relying only on numerical summaries would im taking by faith that the salient features of the relati ship between the variables have been captured. W one or a few numerical summaries are used to cha terize the relationship between two variables, there risk of missing important features or, worse, of be misled.[13] Thus, a scatterplot matrix was used to su marize relationships between covariables (Figure With the weak correlations observed between cov ates, the dimensionality of the covariate vector was reduced; hence, all covariates were used in the sul quent exploratory modeling step.

Structure Revelation and Covariate Selection

NONMEM[14] was used to select a pharmacokin model (i.e., a two-compartment model) that best scribed the data. With this model, post hoc emp Bayesian individual parameter estimates were ge ated and used with the covariates as "data" in the ploratory modeling step for covariate selection data structure revelation. For the purpose of this a cle, the focus is on the development of a popula pharmacokinetic model for clearance (CL).

A density plot (Figure 4) of CL estimates with bandwidth computed using Silverman's rule of thur clearly shows the presence of two subpopulation the data set. The bimodality in the distribution of i vidual CL estimates suggested the presence of subpopulations. Figure 5 is a pairs plot of the pop tion pharmacokinetic data set showing the relations between individual Bayesian CL estimates and sub demographics (covariates). An S-Plus™ function used to produce the content of each scatterplot loess regression curves superimposed to aid the vi perception of the relationship between the varial This figure can easily be appreciated by dividing it two triangular halves—an upper triangle to the r and a lower triangle to the left. For example, visuali the relationship between CL (ordinate) and CRCL scissa) would necessitate using the triangle to the ri It can be observed that CL is related to CRCL by s nonlinear function.

Using the box plot, the relationship between CL age divided into groups of 10 years was examine can be seen in Figure 6a that CL does not appear to function of age group since the CL values are simila tween age groups. Similarly, sex does not influence CL of this drug (Figure 6b). The distribution of CL

PM3000456977

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

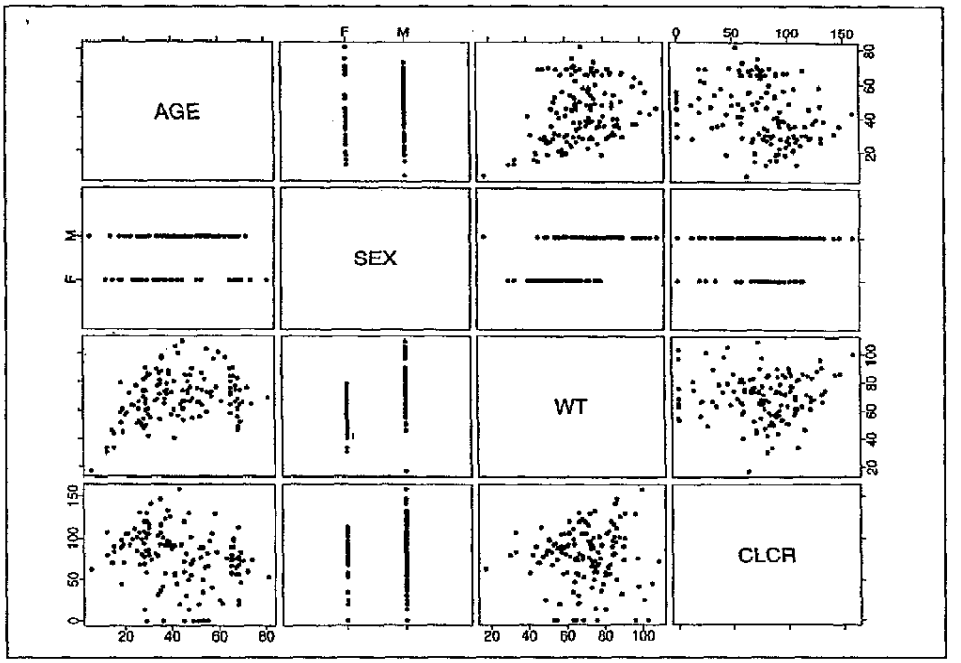


Figure 3. Pairs plot showing relationship between covariates.

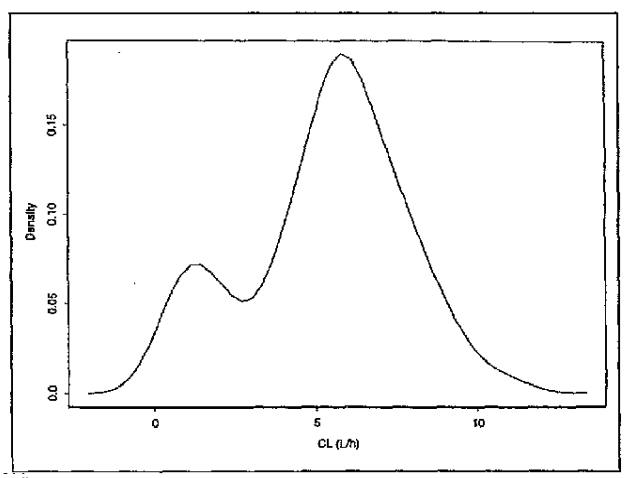


Figure 4. Density plot showing bimodality in the distribution of empiric Bayesian estimates of clearance (CL). The bandwidth for the density plot was determined using Silverman's rule of thumb.[15]

ues between the sexes is similar. Therefore, it can be inferred from these plots that age and sex are probably not predictors of CL.

Both MLR using the general linear modeling (GLM) procedure and GAM[4,8] in S-Plus were used to model the data.

Usually, the multiple linear regression approach is the method of choice for this type of a problem. The multiple linear regression model is expressed in the following form:

$$P = \alpha + \sum_{j=1}^{p} \beta_j X_j + \varepsilon, \quad (1)$$

where $E(\varepsilon) = 0$ and $\text{var}(\varepsilon) = \sigma^2$. This model makes a strong assumption of the linear dependence of E(P) (the expectation of P or mean response) on the predictors. The MLR model is extremely useful and convenient if this assumption holds, even roughly, because it provides a description of the data, summarizes the contribution of each predictor with a single coefficient, and provides a simple method for predicting new observations.

The assumption of linear dependence of the response variable on each of the predictors may not always hold. For many types of data, a change in the mean of the response variable is accompanied by a change in its variance. The GAM approach presents a general perspective for the handling of covariates in a multiple-regression setting. The linear form of $\alpha + \Sigma_j \beta_j X_j$ is replaced with the additive form $\alpha + \Sigma_j f_j(X_j)$, that is,

$$P = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \quad (2)$$

PM3000456978

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

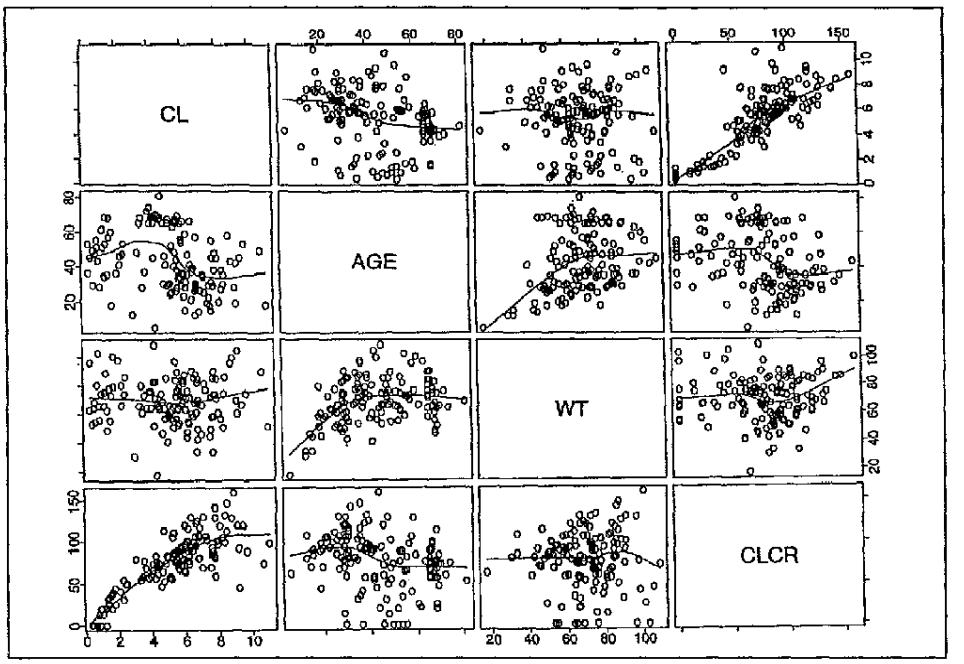


Figure 5. Pairs plot showing relationship between CL and the covariables.

where $f_j(X_j)$ is an arbitrary univariate function that is either a linear function or a smoothing spline. Since each covariate is represented separately in equation (2), GAM retains the important interpretive feature of the linear model: the variation of the fitted response surface holding all but one predictor fixed does not depend on the values of the other predictors. In practice, this means that once the additive model is fitted to the data, one can plot the p-coordinate functions separately to examine the roles of predictors in modeling the response. The estimated function forms of GAM are analogues of the coefficients in multiple linear regression. Thus, separate functions are introduced to allow for nonlinearity and heterogeneous variances. This is closer to a reparameterization of the model than to a reexpression of the response. Generalized additive models are a group of models that are as tractable as the linear model is but do not force the data into unnatural scales.

For both MLR and GAM, the data were subjected to stepwise (single-term addition/deletion) modeling procedure. Each covariate was allowed to enter the model in any of several functional representations. The Akaike information criterion (AIC) was used as the model selection criterion. AIC is a criterion for judging model quality. At each step, the model is changed by addition or deletion of a covariate that results in the largest decrease in AIC. The search was stopped when the AIC reached a minimum value. In the building of the GAM model, each covariate was allowed to enter the model linearly or as a spline function. The scatterplot smoothing function in S-Plus was used.

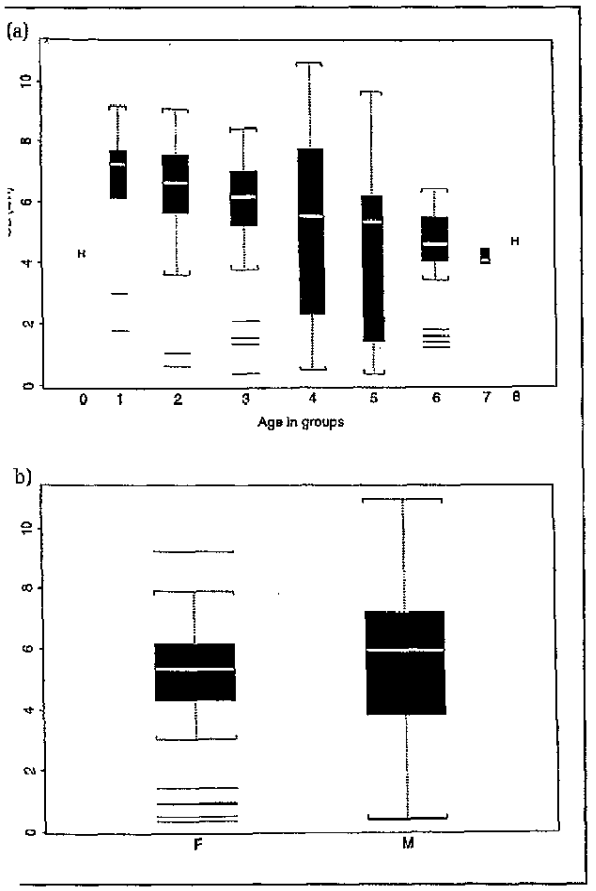
Cook's distance diagnostic was applied to the results of the MLR, and it did not reveal any significant leverage observations in the data. Cook's distance diagnostic measures the effect of the ith individual (observation) on the coefficient vector.

The partial residual plots in Figure 7 summarize the results of the MLR and GAM approaches. The inadequacy of the linear model relating CRCL to CL is obvious with the poor prediction of subjects whose CRCL values were below 55 ml/min (Figure 7a). There are two distinguishable clouds of points above and below the CRCL value of 55 ml/min. The data points above the 55 ml/min CRCL value are scattered randomly around the regression line, while those below are not. The latter are poorly predicted. These two groups of points represent two subpopulations seen in Figure 1b. The partial residual plots reveal the inadequacy of the linear model.

The relationship between CL and age is flat, more or less (Figure 7b). This confirmed the finding from the box plots (Figure 6a) that CL was not a function of age. Both modeling procedures (i.e., MLR and GAM) did not select weight and sex as predictors of CL. The finding that sex was not a predictor of CL served as a confir-

PM3000456979

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001



re 6. *(a) Box plots of the distribution of empiric Bayesian CL estes by age groups. The age groups were created at intervals of 10 s. The minimum age was 5 years and the maximum 81 years. (b) olots of the distribution of empiric Bayesian CL estimates by sex.*

ion of the result of the exploratory graphical analy- with the box plot (Figure 6b).

'he results of the GAM model in Figure 7c show that :L as a nonlinear predictor of CL was better than linear function of the MLR. The detection of nonarity, however, is the central motivation for partial dual plots. Although a nonlinear function (Figure described the data better than the linear function Figure 7a), it was still inadequate. The bulk of the ects with low CRCL values were still poorly pre- ed. This informed the testing in NONMEM of a linear model of CRCL as a predictor of CL and two rate linear models for the two subpopulations as rmed by the distribution of CL and CRCL values the GAM results. The relationship between the ial residuals and age is flat for the majority of the

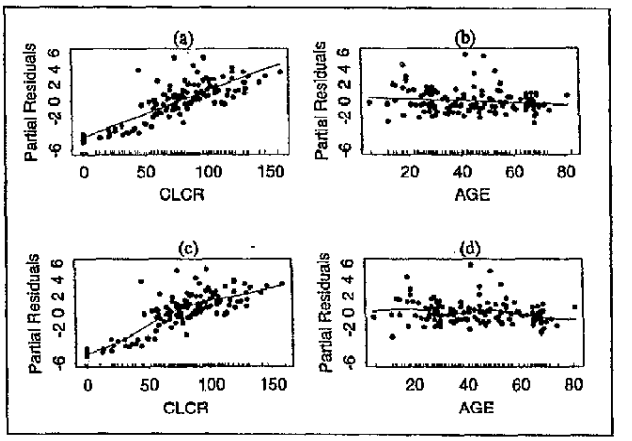


Figure 7. MLR and GAM analyses results: scatterplots of partial residuals of CL (L/h) versus (a) CRCL (ml/min) and (b) age (years) from MLR analysis, CL (L/h) versus (c) CRCL (ml/min), and (d) age (years) from GAM analysis. The same scale is used for the ordinate in each plot so that the relative importance of each covariate can be compared. (Extracted from Ette and Ludden[4])

data. Only a few observations appear to skew the values at extremes of age (Figure 7d).[4]

TBM is an exploratory modeling technique for uncovering structure in the data and assessing the adequacy of linear models.[4,16] It operates only on ranks of the data. This aspect of TBM for a numeric explanatory variable renders it invariant to monotone transformations of the explanatory variable. It automatically incorporates interactions between covariates such as when one parameter-covariate relationship depends on another covariate. These are important advantages that TBM has over GAM. The result of TBM is shown in Figure 8. At each node, every covariate is checked for a possible split point where the variances of the two subsets that would result from the split are most different. That is, the split is carried out at the point where the objective function value (the deviance that in the usual case is just the sum of squares) decreases the most. As implemented in S-Plus, this process is continued in each branch of the split until the terminal contains 10 or less data points. CRCL was the most significant predictor of CL. The numeric value indicated at each split node is the breakpoint value of the covariate used for the split at that node. For instance, 44.8 ml/min CRCL value was used at the root node for the split. The values displayed at the leaves (terminals of the splits) are the average CL values for the homogeneous group of subjects selected into the particular portion of the split. The histograms of the splits for CL and CRCL in the ordinate direction show splits of CL almost matching

PM3000456980

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

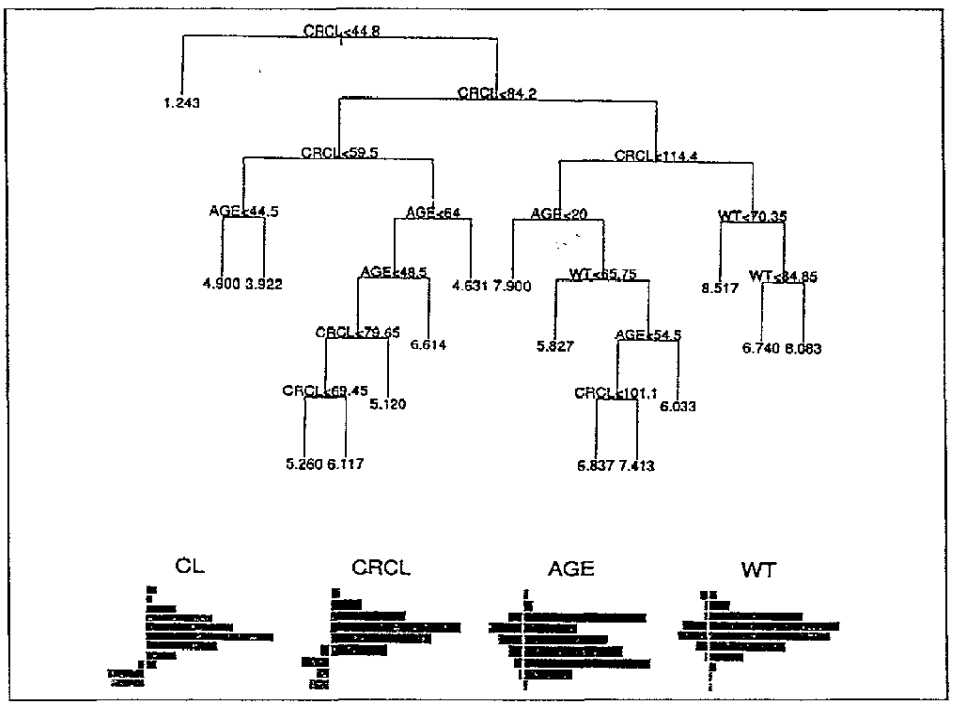


Figure 8. A tree-based model for CL. The top panel displays the labeled dengrogram. The numeric value indicated at each split node is the breakpoint value of the covariate used for the split at that node. The values displayed at the leaves (terminals of the splits) are the average CL values for the homogeneous group of subjects selected into the particular portion of the split. The lower panel displays a side-by-side histogram for CL (L/h), CRCL (ml/min), age (years), and WT (kg). The left-side histogram summarizes the observations following the left split and similarly for the right split. (Extracted from Ette and Ludden[4])

those of CRCL (Figure 8), suggesting two levels of response by two subpopulations.[4]

NONMEM Modeling to Develop the Population Pharmacokinetic Model

The PPK model for CL of this drug was developed in NONMEM by backward elimination using the covariates CRCL and age. The final NONMEM PPK model only used CRCL as a predictor of CL with different slopes and intercepts to characterize the subpopulations.[3] The models estimating CL in the two subpopulations were the following:

Model I: CL, (L/h) = 1.0 (20.0%) + 0.25 (44.0%) • CRCL for individuals with a shallow slope of the CL regression model (subpopulation I).

Model II: CL_2 (L/h) = 1.0 (20.0%) • 0.27 (35.0%) + 0.25 (44.0%) • 4.43 (45.0%) • CRCL for the group of individuals with a steep slope of the CL regression model (subpopulation II).

The data in parentheses represent the percent relative standard error, a measure of precision associated with the estimation of these parameters. The models above were used for recommending doses for patients in the two subpopulations.

The result obtained from NONMEM modeling was predicted from exploratory examination and modeling of the data and confirms the usefulness of exploratory data examination and modeling in the discovery of knowledge from a large pharmacokinetic data set.

Reliability of Estimates

The reliability of the parameter estimates was checked using a nonparametric technique—the jackknife technique.[4] The nonlinearity of a statistical model and ill conditioning of a given problem can produce numerical difficulties and force an estimation algorithm into a false minimum.

The preciseness of the primary parameters can be estimated from the final fit of the multiexponential function to the data, but they are of doubtful validity if the model is severely nonlinear.[17] The preciseness of the secondary parameters (in this case, variability) is likely to be even less reliable. Consequently, the results of statistical tests carried out with preciseness estimated from the final fit could easily be misleading—thus the need to assess estimates' reliability. The first-order method in NONMEM, which was used for the development of the population pharmacokinetic model, can yield estimates of parameters that are sometimes bi-

PM3000456981

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

Table I Comparison of Normal Theory Parameter and Variance Estimates with Jackknife Estimators

	Normal Theory			Jackknife[a]		
Parameter	Estimate	*SE*	%RSE	Estimate	*SE*	%JKK *SE*
θ_1	0.990	0.200	20.200	0.780	0.075	9.615
θ_2	0.270	0.090	33.330	0.910	0.335	36.813
θ_3	0.250	0.110	44.000	0.260	0.055	21.150
θ_4	4.430	2.000	45.460	3.513	0.347	9.787

a. A total of 10% of subjects were removed/run. θ_1 is the intercept for the regression of CRCL as a predictor of CL (i.e., CL_1) in subpopulation I. θ_2 is the intercept scaling factor for the regression of CL (i.e., CL_2) on CRCL in subpopulation II. θ_3 is the regression coefficient for CRCL as a predictor of CL in subpopulation I. θ_4 is the regression coefficient scaling factor for CRCL as a predictor of CL in subpopulation II. (Excerpted from Ette and Ludden[4])

ased. A possible way of reducing bias in parameter estimates and of calculating realistic variances for them is to subject the data to the jackknife technique.[18-20] The technique requires little by way of assumption or analysis. A naive Student *t* approximation for the standardized jackknife estimator[19] was used in the investigation.

Most of the jackknife estimators for regression coefficients and variability were similar to the NONMEM estimates (Table I). In some cases, these were more precise than the NONMEM estimates. The similarity of most of the final NONMEM parameter estimates to jackknife estimators indicated that the NONMEM estimates were relatively unbiased and precise. The difference between the normal theory and jackknife estimates of θ_2 was a consequence of the parameterization of the NONMEM model for CL. This parameter was used as the intercept scaling parameter for subpopulation II and, therefore, sensitive to changes in the jackknife data sets when some subjects from subpopulation I (the smaller subpopulation) were not included in some of the JKK data sets. This is a limitation of the jackknife technique. The population pharmacokinetic estimates obtained with the final (irreducible) NONMEM model were, therefore, found to be reliable.

The reliability of parameter estimates can also be determined using the bootstrap approach (see Ette,[21] for instance, for the application of the bootstrap to population pharmacokinetic modeling). The use of the bootstrap for the determination of the reliability of parameter estimates is outside the scope of this article.

Communication of Discovered Knowledge

The population models developed for estimating clearance in the two subpopulations were used in designing two dosage guidelines for patients in the respective subpopulations. Patients with CRCL less than 50 ml/min (subpopulation I) had their drug clearance values estimated with model I, while patients with CRCL greater than or equal to 50 ml/min had their drug clearance values estimated with model II.

DISCUSSION

In the example discussed, a hypothesis of bimodality in CL distribution was developed, which was further investigated and confirmed. In fact, this hypothesis formed the framework for the development of the population pharmacokinetic model. The population pharmacokinetic model developed was used for designing dosage guidelines for the drug. It could also be used in clinical trial simulation for the planning of a subsequent clinical trial with the drug. These, in essence, are avenues for applying the discovered pharmacokinetic knowledge.

The use of population modeling for knowledge discovery from a pharmacokinetic/pharmacodynamic data set accumulated from dose-finding studies addresses the shortcoming of traditional ways of analyzing such data through the discovery of interesting patterns. The shortcoming of traditional analysis of pharmacokinetic/pharmacodynamic data is that one does not know which questions to ask (hypothesis to test) to discover hidden knowledge.

The process of knowledge discovery from a population pharmacokinetic data set (i.e., from data assembly to population pharmacokinetic model development) is an iterative and heterogeneous task, and it might be argued that it is difficult to summarize the process in terms of the formalized steps described in this article. However, over the past 20 years, considerable knowledge has been accumulated on this subject; therefore, it has become necessary to delineate the steps taken in the process of knowledge discovery from a population pharmacokinetic data set.

PM3000456982

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001

CONCLUSION

A structured approach for knowledge discovery from a large pharmacokinetic data set has been described. The approach described in this article lays out systematically how hidden knowledge can be discovered from a population pharmacokinetic data set with the use of modern graphical, modeling, and statistical approaches. The use of these techniques for knowledge discovery from a large pharmacokinetic data set gives the data analyst the liberty to choose statistical methodology appropriate to the problem at hand with the maximization of information extraction, rather than on the basis of mathematical/statistical tractability.

REFERENCES

1. Aarons L: Population pharmacokinetics: theory and practice. *Br J Clin Pharmacol* 1991;32:669-670.

2. Fayyad U, Piatetsky-Shapiro G, Smyth P: From data mining to knowledge discovery: an overview, in: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthursamy R (eds.), *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: Calit AIII Press/MIT Press, 1996;1-34.

3. S-PLUS, version 3.4. Seattle, WA: Statistical Sciences, Inc., 1995.

4. Ette EI, Ludden TM: Population pharmacokinetic modeling: the importance of informative graphics. *Pharm Res* 1995;12:1845-1855.

5. Fellegi IP: Data, statistics, information: some issues of the Canadian social statistics scene. *J Official Statistics* 1988;4:53-65.

6. *Guidance on Population Pharmacokinetics*. Rockville, MD: U.S. Food and Drug Administration, Department of Health and Human Services, 1999.

7. Grasela TH, Antal EJ, Fiedler-Kelley J, et al: An automated drug concentration screening and quality assurance program for clinical trials. *Drug Info J* 1999;33:273-279.

8. Hastie TJ: Generalized additive models, in: Chambers JM, Hastie TJ (eds.), *Statistical Models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1992;249-307.

9. Shortliffe E, Perreault LE (eds.): *Medical Informatics: Computer Applications in Health Care*. Reading, MA: Addison-Wesley, 1990.

10. Rubin DB: Inference and missing data. *Biometrika* 1976;63: 581-582.

11. Rubin DB: Multiple imputation after 18+ years. *J Am Stat Assoc* 1996;91:473-489.

12. Shapiro SS, Francia RS: An approximate analysis of variance test for normality. *J Am Stat Assoc* 1972;67:215-216.

13. Ette EI: Statistical graphics in pharmacokinetics and pharmacodynamics: a tutorial. *Ann Pharmacother* 1998;32:818-828.

14. Beal SL, Sheiner LB: *NONMEM Users' Guide*, parts I-VI. San Francisco: Division of Clinical Pharmacology, University of San Francisco, 1979-1992.

15. Silverman BW: *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.

16. Brieman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.

17. Drapper NR, Smith H: *Applied Regression Analysis*. New York: John Wiley, 1966;263-304.

18. Quenouille MH: Notes on bias estimation. *Biometrika* 1956;43: 353-360.

19. Miller RG: The jackknife: a review. *Biometrika* 1974;61:1-15.

20. Hinkley DV: Jackknife confidence limits using Student t approximations. *Biometrika* 1977;64:21-28.

21. Ette EI: Population model stability and performance. *J Clin Pharmacol* 1997;37:486-495.

PM3000456983

Source: https://www.industrydocuments.ucsf.edu/docs/mmcl0001